AVOIDING DUPLICATIONS IN PERSON DETECTION ACROSS VIDEO FRAMES

Soujanya Pradheepa Lohanathen

238040P

Master of Science(Major Component of Research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July 2025

AVOIDING DUPLICATIONS IN PERSON DETECTION ACROSS VIDEO FRAMES

Soujanya Pradheepa Lohanathen

238040P

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Science(Major Component of Research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

July 2025

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 02.08.2025

The above candidate has carried out research for the Master of Science (Major component of Research)Thesis under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. Chandana Gamage

Signature of the Supervisor: Date: 02.08.2025

Name of Supervisor: Dr. Sulochana Sooriyaarachchi

Signature of the Supervisor: Date: 02.08.2025

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support, guidance, and encouragement of many individuals and institutions to whom I am deeply grateful.

First and foremost, I offer my heartfelt thanks to my supervisors, Prof. Chandana Gamage and Dr. Sulochana Sooriyaarachchi. Their visionary insights and meticulous attention provided the perfect balance of breadth and depth throughout this work. Their patient mentoring, constructive criticism, and unwavering faith in my abilities motivated me to overcome challenges and refine my ideas into a cohesive, robust system.

I am also indebted to my examiners, Prof. Chathura de Silva and Dr. Kutila Gunasekara. Their careful review of my research during progress reviews and their probing questions led me to strengthen my arguments and clarify my contributions. Their expertise and thoughtful suggestions have elevated the quality of this research.

I would like to thank Dr. Uthayashankar Thayasivam, Head of the Department of Computer Science and Engineering, University of Moratuwa. I extend my gratitude to all faculty members and administrative staff, whose efficient handling of academic and logistical matters allowed me to focus wholly on my project.

I owe a special debt of gratitude to Dr. Sanka Rasnayake of the National University of Singapore. His willingness to discuss ideas across time zones and his candid feedback on preliminary results were invaluable, and his encouragement inspired me to push the boundaries of my work.

Within the IntelliSense Lab, I found a community of colleagues who generously shared their technical know-how and moral support. I especially thank my lab mates for brainstorming with me during early experiments and for the camaraderie that made long hours in the lab both productive and enjoyable.

On a personal level, I am profoundly grateful to my parents and siblings. Their belief in my potential has been my greatest source of strength. To my friends—who lent a listening ear and offered words of encouragement—I extend my deepest thanks.

Finally, I would like to acknowledge all those who directly or indirectly contributed to this research: the organizers of open-source datasets and tools, the peer reviewers whose work informed my literature review, and the wider academic community whose discoveries laid the groundwork for this thesis. Your collective efforts have made this journey possible, and I am honoured to add my contributions to the field.

ABSTRACT

Person re-identification (Re-ID) is a cornerstone of modern video surveillance and smart-city applications, demanding the reliable matching of pedestrian images across non-overlapping cameras despite variations in pose, lighting, background clutter, and occlusion. Here, a person re-identification (Re-ID) system built around a ResNet-50 backbone augmented with multi-level attention and part-aware Transformer encoding is presented. The network begins by extracting deep feature maps from pedestrian images, which are then refined through a channel-wise squeeze-and-excitation block and a spatial attention module: together, these attentional layers suppress background clutter and highlight discriminative cues—such as clothing textures and carried objects—by adaptively weighting feature dimensions and spatial locations. To capture structural dependencies across body regions, the attention-refined feature map is partitioned into horizontal strips corresponding to semantic parts (head, torso, legs), each of which is fed into a lightweight Transformer encoder that dynamically models inter-part relationships, enabling robust identification under pose variation and partial occlusion.

Training is stabilized and accelerated via mixed-precision optimization with automatic gradient scaling and gradient clipping, alongside a label-smoothed cross-entropy loss that mitigates overconfidence. A two-stage learning-rate schedule—a brief linear warm-up followed by cosine-annealing decay—ensures rapid initial convergence without catastrophic divergence. At inference, global descriptors are efficiently extracted and pairwise distances computed to evaluate mean average precision (mAP) and Rank-1 accuracy on the Market-1501 benchmark.

Empirical results demonstrate that this architecture achieves competitive retrieval performance—regularly exceeding 0.74 mAP and 0.90 Rank-1 accuracy while maintaining computational efficiency and ease of extension. All data-processing pipelines, training scripts, and evaluation code are fully open-source, providing a reproducible framework for future advances in attention-driven person Re-ID.

Keywords: Person Re-identification, Unique person counting, Video processing, Surveillance applications

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor i

Abstract iii

Table of Contents iv

List of Figures vii

List of Tables viii

1 Introduction 1

1.1 Problem Definition and Challenges 1

1.2 Deep Learning for Person Re-Identification 2

1.3 Transformers and Inter-Part Relationship Modelling 2

1.4 Real-Time Video-Based Re-Identification 3

1.5 Main Contributions Of The Research 3

1.6 Thesis Organization 4

2 Background & Related Work 5

2.1 Introduction 5

2.2 Foundations of Person Re-Identification 5

2.2.1 Feature Engineering and Metric Learning 5

2.2.2 Multi-Frame and Sequence Approaches 5

2.3 Deep Learning: The Modern Paradigm 6

2.3.1 CNN and RNN Architectures 6

2.3.2 Weakly and Few-Shot Supervised Approaches 6

2.4 Attention Mechanisms in Person Re-ID 6

2.4.1 Motivation and General Principles 6

2.4.2 Spatial Attention 7

2.4.3 Temporal Attention 7

2.4.4 Joint Spatial-Temporal Attention 7

2.4.5 Progressive and Hierarchical Attention 8

2.5 Part-Based Modelling 8

2.5.1 Motivation and Key Paradigms 8
2.5.2 Horizontal Partitioning 8
2.5.3 Pose-Guided and Graph-Based Models 8
2.5.4 Adaptive and Attention-Based Part Fusion 9
2.6 Transformer-based Architectures 9
2.6.1 Vision Transformers: Motivation and Properties 9
2.6.2 Transformers in Video-based Re-ID 9
2.6.3 Overcoming Limitations: Information Loss, Fragmentation, and Modality Gaps 10
2.6.4 Attribute-Enhanced and Multi granularity Transformers 10
2.7 Challenges: Occlusion, Open-World Settings, and Real-Time 10
3 Methodology 12
3.1 Data Preprocessing and Augmentation 12
3.1.1 Loading 12
3.1.2 Resizing 12
3.1.3 Data Augmentation 13
3.1.4 Normalization 13
3.1.5 Implementation Details 13
3.2 Model Components and Architectural Design 14
3.2.1 Backbone Feature Extraction 14
3.2.2 Channel Attention Module 17
3.2.3 Spatial Attention Module 19
3.2.4 Part-Based Feature Extraction 22
3.2.5 Part-Aware Transformer Encoding 23
3.3 Descriptor Aggregation and Classification Head 26
3.3.1 Concatenation and Batch Normalization 27
3.3.2 L2-Normalization 28
3.3.3 Classification Layer for Training 28
3.4 Loss Functions 29
3.4.1 Label-Smoothed Cross-Entropy 29
3.4.2 Batch-Hard Triplet Loss 29

3.4.3 Total Loss 30
3.5 Training Protocol 30
3.5.1 Optimizer and Mixed Precision 30
3.5.2 Learning-Rate Schedule 30
3.5.3 Training Loop 30
3.5.4 Inference and Retrieval Pipeline 31
3.5.5 Optional Re-Ranking 32
3.5.6 Complexity Analysis 33
4 Experiments and Results 34
4.1 Experimental Setup 34
4.1.1 Dataset Description 34
4.1.2 Evaluation Metrics 34
4.1.3 Implementation Details 34
4.2 Baseline Performance 35
4.3 Ablation Studies 35
4.3.1 Component Ablation 36
4.3.2 Effect of Partition Count 37
4.3.3 Convergence and Stability 37
4.3.4 Qualitative Retrieval Examples 38
4.3.5 Comparison with State of the Art 39
4.4 Summary 39
5 Conclusion and Future Work 40
5.1 Conclusion 40
5.2 Limitations 41
5.3 Future Work 42
5.3.1 Unsupervised Domain Adaptation 42
5.3.2 Dynamic Part Partitioning 43
5.3.3 Temporal and Video-Based Modelling 43
5.3.4 Multi-Modal and Infrared Fusion 43
5.3.5 Lightweight Model Compression 44
5.3.6 Explainability and Human-In-the-Loop Learning 44

LIST OF FIGURES

Figure	Description	Page
Figure 3.1	Data Preprocessing Pipeline: Load $\rightarrow$ Resize $\rightarrow$ Augment $\rightarrow$ Normalize.	12
Figure 3.2	System Architecture Flowchart: Attention-Enhanced Part-Aware Re-ID Pipeline	15
Figure 3.3	ResNet-50 backbone module with downsampling to $2048 \times 8 \times 4$.	16
Figure 3.4	Structure of the Channel Attention Module using a Squeeze-and-Excitation (SE) block	18
Figure 3.5	The concatenation results in a tensor of shape $[B, 6 \times 2048]$ where B is the batch size. The attention maps are applied via element-wise multiplication to the feature maps before concatenation, enhancing discriminative regions in each part. [30]	20
Figure 3.6	Horizontal partitioning of the attention-refined feature map into $P = 6$ stripes along the vertical axis	22
Figure 3.7	Part-Aware Transformer architecture for modelling inter-part relationships	24
Figure 3.8	Descriptor aggregation and classification pipeline	27
Figure 3.9	Combined loss function used for training	29
Figure 4.1	Comparison of the accuracy with different partition values	37
Figure 4.2	mAP curves	38
Figure 4.3	Qualitative results	38

LIST OF TABLES

Table	Description	Page
Table 2.1	Comparison of Person Re-ID Approaches	11
Table 4.1	Baseline performance (ResNet-50 + global pooling + CE)	35
Table 4.2	Ablation results on Market-1501	36
Table 4.3	Qualitative retrieval results for three query images and their Top-5 matches.	39
Table 4.4	Comparison to recent state-of-the-art on Market-1501	39